ANALIZA STATISTICĂ A DATELOR ECONOMICE PRIN ALGORITMI DATA MINING DE ARBORI DE DECIZIE

LAVINIU AURELIAN BĂDULESCU
Universitatea din Craiova; Facultatea de Automatică, Calculatoare şi Electronică, Secţia Inginerie Software; Craiova, str. G. Fotino , nr. 2, bl. b8, 3, 11, 200481, Dolj; mobil 0722276554;
E-mail: *laviniu_aurelian_badulescu@yahoo.com*

In contrast with statistical analysis, data mining analyzes all the relevant data in database and extracts hidden patterns. Data Mining Decision Tree Algorithms generate classification or estimation models. The algorithms use a splitting criterion to determine the most predictive factor and place it as the first decision point in the tree. Specific decision tree include CART and CHAID.

1. INTRODUCERE

Data Mining este în anumite privinţe o extensie a statisticii la care au fost adăugate elemente de inteligenţă artificială şi Machine Learning. Ca şi statistica, Data Mining nu este o soluţie de afaceri, ea este doar o tehnologie [8]. În contrast cu analiza statistică, Data Mining analizează toate datele relevante din baza de date şi extrage modelele (*patterns*) ascunse [2]. Metodele de lucru ale statisticii sunt combinate cu metodele specifice domeniului Machine Learning şi sunt ajustate pentru estimarea modelului optim de performanţă din bazele de date [6].

Un arbore de decizie (*Decision Tree*) este un model de clasificare sau estimare care poate fi privit ca un arbore. Fiecare subarbore din componenţa sa reprezintă un răspuns la o întrebare de clasificare, frunzele arborelui sunt partiţii sau segmentări ale setului de date în funcţie de clasificarea realizată, iar nodurile prezintă informaţii statistice.

Ideea de bază a algoritmilor de arbori de decizie este utilizarea unui criteriu de divizare pentru a determina cel mai predictiv factor şi amplasarea lui ca prim punct de decizie în arbore şi în continuare să execute o căutare de factori predictivi pentru a construi subarborii până când nu mai există date de procesat. Reducerea arborelui (*tree pruning*) creşte acurateţea la datele-zgomot şi poate fi realizată atunci când arborele este în construcţie (*pre-pruning*), sau după construcţia arborelui (*post-pruning*).[5]

Arborele de decizie generează un output cu o interpretare uşoară pentru marketing şi o identificare facilă a variabilelor semnificative în luarea deciziilor manageriale. La construirea modelului arborescent se pot utiliza variabilele originale care nu au fost transformate sau normalizate. Modelul de arbore de decizie va crea reguli asupra datelor de estimat variabila ţintă. Metodele specifice de arbore de decizie includ arborii de clasificare şi regresie (*Classification and Regression Trees*: CART) şi detecţia automată a interacţiunii χ^2 (*Chi Square Automatic Interaction Detection*: CHAID). Ei furnizează o mulţime de reguli ce pot fi aplicate pentru un set de date neclasificat, pentru a estima care înregistrări

vor avea o anumită ieşire. CART segmentează un set de date creând subarbori binari, în timp ce CHAID segmentează setul de date creând subarbori oarecare, utilizând teste χ^2. CART necesită de obicei mai puțină pregătire a datelor decât CHAID.[1]

Deşi arborii de decizie au fost dezvoltați inițial ca instrumente exploratorii pentru rafinarea şi preprocesarea datelor pentru tehnici statistice, cum ar fi regresia logică, ei sunt din ce în ce mai mult utilizați pentru predicție. Analiza regresiei este o tehnică statistică tradițională pentru găsirea unei funcții care descrie relația dintre un număr de variabile şi o valoare care se doreşte estimată. Această tehnică utilizează, în general, intrări numerice. De obicei, este necesară o preprocesare. Cele mai utilizate tehnici de regresie sunt: *regresia polinomială* (*polynomial regression*)**,** extensie a regresiei lineare şi *regresia logică* (*logistic regression*), ieşirea în acest caz fiind 1 sau 0 [3]. Arborii de decizie luați în considerare în problema analizelor de regresie sunt numiți arbori de regresie.[9]

2. ALGORITMII CART ŞI CHAID

Algoritmul CART este un algoritm de explorare şi predicție[4] care alege fiecare predictor la construirea arborelui astfel încât să scadă dezordinea datelor. Măsura pe baza căreia este preferat un predictor altuia este valoarea entropiei. Algoritmul CART este relativ robust în raport cu datele lipsă. Dacă o valoare lipseşte pentru un predictor particular într-o înregistrare particulară, la construirea arborelui acea înregistrare nu va fi utilizată în realizarea determinării ramificării optimale. Când CART este utilizat pentru a prezice asupra unor date noi, valorile lipsă pot fi manipulate prin intermediul substitutelor (*surrogates*). Substitutele sunt valori de ramificare şi predictori care simulează ramificarea reală din arbore şi pot fi utilizate când lipsesc datele pentru predictorul dorit. De exemplu, deşi mărimea la pantofi nu este un predictor perfect pentru înălțimea unei persoane, ea poate fi folosită ca un substitut în încercarea de a simula o ramificare bazată pe înălțime când acea informație lipseşte dintr-o înregistrare particulară ce trebuie utilizată în estimarea cu modelul CART.

CHAID diferă de CART în modul cum alege ramificarea. Pentru alegerea ramificării optimale, CHAID se bazează pe testul χ^2 din tabelele de contingență pentru a determina care predictor categorial este cel mai departe de independență cu valorile estimate. Algoritmul CHAID este popular în cercetările de marketing în contextul studiilor de segmentare a pieței. Putând fi utilizați atât pentru predicție cât şi pentru clasificare, algoritmii CART şi CHAID pot fi aplicați pentru analiza problemelor de tip regresie sau de tip clasificare. Prezentăm paşii urmați la dezvoltarea algoritmului CHAID.

1. Pregătirea predictorilor. Se construiesc predictorii categoriali din predictorii continui prin împărțirea distribuției continue într-un număr de categorii cu un număr aproximativ egal de observații. Pentru predictorii categoriali, categoriile (clasele) sunt definite de la sine.

2. Fuziunea categoriilor. Parcurgem repetat predictorii pentru a determina pentru fiecare predictor perechea de categorii predictor care sunt cel mai puțin semnificative în raport cu variabila dependentă; pentru problemele de clasificare (unde variabila dependentă este categorială), se va evalua un test χ^2 (Pearson χ^2); pentru problemele de regresie (unde variabila dependentă este continuă), se va evalua un test F. Dacă testul respectiv pentru o pereche dată de categorii predictor nu este semnificant statistic, atunci se vor fuziona categoriile predictor respective şi se va repeta acest pas (*i.e.* se va găsi următoarea pereche

de categorii, care acum pot include categoriile anterior fuzionate). Dacă perechea de categorii predictor este semnificativă statistic, atunci se va estima un test Bonferroni *p*-valoare ajustată pentru mulțimea categoriilor predictorului respectiv.

3. Selectarea valorii de separare. Alegem variabila predictor de separare cu cea mai mică *p*-valoare ajustată, adică variabila predictor care produce cea mai semnificativă separare; dacă cea mai mică *p*-valoare ajustată (Bonferroni) pentru orice predictor este mai mare decât o anumită valoare de separare α, atunci nu va mai fi executată nici o separare şi nodul respectiv este o frunză.
Acest proces continuă până când nu mai poate fi realizată nici o separare.[7]

3. CONCLUZII

Arborele de decizie şi algoritmul care îl creează pot fi complicați, însă rezultatul poate fi prezentat într-un mod uşor de înțeles, lucru care poate fi extrem de folositor în luarea deciziilor în afaceri. Astfel arborele de decizie este situat în topul modelelor predictive. El poate fi utilizat însă, în egală măsură, şi în aplicațiile de clasificare ce sunt solicitate în diverse domenii cum ar fi experimentele ştiințifice, aprobările de credite, target marketing, store location, analizele financiare, customer segmentation, detectarea fraudelor etc.

Să observăm două elemente interesante la acest tip de arbore:

- el divide datele la fiecare punct de ramificare fără să piardă nici o dată, numărul total de înregistrări din nodul părinte fiind egal cu suma înregistrărilor conținute în cei doi subarbori fii;
- este uşor de înțeles cum a fost construit modelul, în contrast cu alte modelele concurente cum ar fi rețele neuronale etc.

Datorită înaltului lor nivel de automatism şi uşurinței de translatare a modelelor construite cu arbori de decizie în SQL, pentru utilizarea în baze de date relaționale, tehnologia este uşor de integrat în procese IT deja existentele, necesitând puțină preprocesare şi reducere a datelor, sau extragere a lor cu scop precis pentru Data Mining.

BIBLIOGRAFIE

1. ***, “An Overview of Data Mining at Dun & Bradstreet”, Data Intelligence Group White Paper 95/01, 1995.
2. Baragoin, C., Andersen, C., M., Bayerl, S., Bent, G., Lee, J., Schommer, C., “Mining Your Own Business in Retail Using DB2 Intelligent Miner for Data”, International Technical Support Organization, International Business Machines Corporation, RedBook, San Jose, California, 2001, p. 21.
3. Baragoin, C., Andersen, C., M., Bayerl, S., Bent, G., Lee, J., Schommer, C., “Mining Your Own Business in Telecoms Using DB2 Intelligent Miner for Data”, International Technical Support Organization, International Business Machines Corporation, RedBook, San Jose, California, 2001.
4. Breiman, L., Friedman, J., Olshen, R., Stone, C., “Classification and Regression Trees”, Stanford University and the University of California, Berkeley, 1984.
5. Nepomnjashiy, A., “Data Mining Algorithms: Microsoft SQL Server 2000 vs. "Yukon" SQL Server”, DatabaseJournal.com, 2004, http://www.databasejournal.com/.
6. Ratner, B., “Statistical Modeling and Analysis for Database Marketing: Effective Techniques for Mining Big Data”, Chapman & Hall/CRC, 2003.